HWG | HARRIS, WILTSHIRE & GRANNIS LLP

December 21, 2017

Ex Parte

Mr. David Schmidt
TRS Fund Program Coordinator
Office of Managing Director
Federal Communications Commission
445 12th Street, SW
Washington, DC 20554

Re: *Telecommunications Relay Services and Speech-to-Speech Services for Individuals with Hearing and Speech Disabilities*, CG Docket No. 03-123; *Misuse of Internet Protocol (IP) Captioned Telephone Service*, CG Docket No. 13-24

Dear Mr. Schmidt:

CaptionCall is moving forward with the development of a fully automated speech recognition ("ASR") product and applauds the Commission for engaging MITRE to begin the process of examining the potential benefits and limitations of ASR. We further encourage the Commission to seek peer review of MITRE's test plans and, later, results.

Attached is a letter that aims to constructively assist the Commission and MITRE in examining and evaluating ASR. We previously shared a substantially similar letter with MITRE.

Sincerely,

John T. Nakahata
Counsel to CaptionCall, LLC and Sorenson Communications LLC

cc: Nicholas Degani
Claude Aiken
Zenji Nakazawa
Amy Bender
Travis Litman
Nathan Eagan
Karen Peltz Strauss
Robert Aldrich
Eliot Greenwald
Michael Scott
Susan Bahr

Attachment 1

CaptionCall Commentary on CAMH Phase 3 Baseline Usability Test Plan

Introduction

MITRE has been undertaking testing of both commercial IP CTS and fully automated speech recognition. This early-round testing understandably has been subject to numerous constraints. In the spirit of constructive and transparent evaluation, this document provides: (1) Observations about to date MITRE's Phase 1 and 2 testing of various ASR systems; (2) Comments on the planned Phase 3 tests; (3) Suggestions for further factors to be included and taken into account; and (4) Suggestions for process improvements including peer review. CaptionCall applauds the FCC's efforts to examine and understand the potential benefits and limitations of new technologies that may improve relay services. We further encourage the Commission to seek peer review of MITRE's test plans and, later, results.

1. Observations regarding MITRE testing to date

We note, as initial tests, Phases 1 and 2 have been systemically constrained as follows:

A. **Only three call types have been tested, all from recordings, and in very small quantities.**
   - Limited Subject Matter and Variability. These were a "doctor's appointment," "information security," and "pizza order." All were adult male voices. To maintain uniformity, the test calls were recordings, rather than conversations between individuals occurring in real time. As such, these do not approximate the flowing, spontaneous, highly variable conversations that in reality occur every day among actual IP CTS users.
   - Audio Signal Quality. Quality of the audio signal was not subject to the variability of real-world environments in which providers operate.
   - Limited Set of Calls. MITRE tested a total of 650 calls across all providers. This means that scale and reliability sufficient to meet both user and regulatory expectations were not tested.

B. **Accuracy and usability measures captured some, but not all, aspects of actual user understanding or comprehension.** In attempting to gauge the accuracy of text streams resulting from the test call transcriptions, the error-rate calculations used only spelling of words and missing words. Punctuation, dysfluencies, prosody, and information transfer that aid readability and comprehension were not considered. Further testing is needed to determine what characteristics are necessary for user comprehension, given the age and demographics of the user population and different levels of residual hearing. In addition, the following merit further examination of the results and calculations, or further clarification and explanation:
   - *The scoring method for Accuracy and Completeness, as described in the Test Plan[1], is confusing and may provide misleading results; in some cases errors potentially could be double counted.* Rather than using a single metric such as Word Error Rate (WER), but rather "Accuracy" and "Completeness" – a count of words in resulting transcriptions. MITRE's report indicates, "Omitted words are reflected in both the accuracy and completeness metrics,"[2] which suggests that Accuracy and Completeness results may be mingled."[3] If Accuracy and Completeness metrics are somehow combined, the resulting scores will be artificially depressed through double-counting of errors. To clarify both method and results, the raw test data and scoring calculation details should be made available for review by providers and other interested parties.
   - *Potential deficiencies in the test plan and execution.* The reports provide no explanation of why all systems were not tested on all conversations, nor any reason for why conversations with

[1] Phase 3 Test Plan, Section 2.3.6, page 6
[2] Phase 3 Test Plan, page 6 and Summary of Phase 2 Usability Testing Results, page 16
[3] Phase 3 Test Plan, page 6 and Summary of Phase 2 Usability Testing Results, page 16

white noise added were excluded from accuracy averages. The reports also do not explain how MITRE derived score averages from individual call results. The opacity of the methodology makes the results hard to understand and of questionable value in comparing among providers or as a reflection of real world IP CTS use cases.

- *Assessments of the ultimate utility of IP CTS call transcriptions to the deaf or hard-of-hearing user may not have employed appropriate methods*. The System Usability Scale (SUS) method's ten questions were designed for testing usability of industrial or business systems and applications (*e.g.*, industrial equipment or office software) rather than consumer services. Thus, the conclusion (page 17) that users do not believe that IP CTS *transcripts* have a high degree "usability" lacks a sound foundation. Moreover, the summary of phases 1 and 2 do not provide details on responses to the SUS questions or subsequent test subject comments, so one cannot assess the utility of the scores themselves with respect to consumer usability. Finally, SUS itself is not an appropriate method for evaluating usability of IP CTS transcriptions, which attempt to improve user comprehension in a context-rich environment. To quote SUS creator John Brooke "*Usability does not exist in any absolute sense; it can only be defined with reference to particular contexts. This, in turn, means that there are no absolute measures of usability..."*[4] Brooke himself characterized SUS as a "quick and dirty"[5] method.

C. **User group samples sizes were extremely small and hence non-representative.** Section 1.2 of Appendix C in the Test Plan indicates that MITRE used a small sample of twenty individuals for "usability" assessment. The summary of Phase 2 testing, section 2.2 notes the use of five participants in Phase 1 and eleven in Phase 2.[6] The use of very small groups may be understandable, but means that the results are at best directional at a very high level.

Based on the foregoing, the FCC should cautiously evaluate and weigh the results of the test and CAMH recommendations before relying on them other than to help formulate further testing and research. It would be preferable for further testing to have an approach that uses the huge volumes of data available from *actual* calls among IP CTS users that occur every day, with call types, sound, and speakers that reflect actual IP CTS usage. Such an approach would be more likely to provide a statistically valid basis for decision making by policy makers.

2. Phase 3 proposed testing

In the Phase 3 test plan indicates several further factors that could limit the ability to draw generalizable conclusions:

A. **The smart phone as test platform is not representative of the population most in need of IP CTS functionality.** Among the more than 26 million U.S. citizens over age 70, just 32 percent have a smart phone, so the majority, more than 17 million, do not.[7] Among this demographic, landline PSTN service is substantially more prevalent in this population at 76 percent,[8] and more frequently relied upon for day-to-day use than smart phones. Further, for this population, smart phones tend to present some basic conceptual, technological, and physical barriers including operating system interfaces and controls and touchscreen "glass keypads," that are a physical challenge to those with vascular circulation problems or impaired motor control. Testing should focus on platforms and application designs likely to be found among those needing IP CTS services to use a telephone with success.

[4] http://uxpajournal.org/wp-content/uploads/pdf/JUS_Brooke_February_2013.pdf
[5] https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html
[6] Summary of Phase 2 Usability Testing Results, page 4
[7] https://www.census.gov/prod/cen2010/briefs/c2010br-09.pdf and http://www.pewinternet.org/2017/05/17/technology-use-among-seniors/
[8] https://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201705.pdf

B. **VoIP applications inherently favor fully automated ASR system performance.** Applications of the type listed among the thirteen candidate applications considered, and the four chosen for testing, are VoIP "soft phone" applications. Such applications favor fully automated ASR in that the transmitted and received audio streams are isolated from each other, masking the effects of hybrid circuit echo and the ragged audio overlaps between calling parties common on analog telephone lines. Conversely, VoIP soft phone applications are over-the-top systems subject to the vagaries of network, wireless and even wired internet service performance. Packet management and loss can produce artifacts and dropouts in conversation that can be as problematic as poor analog service, especially in rural or geographically problematic environments.

C. **Specification of a single, mobile, smart phone device platform.** The Google Pixel phone is not representative of what IP CTS users (who are mostly elderly and on a fixed income) are likely to have or be comfortable using. The retail cost of the 5.5-inch screen Pixel XL with 32GB of storage is nearly $800[9], and likely beyond the financial reach of a population where the median income is about $38,000 per year.[10]

D. **In addition to speech recognition delivery, acquisition and setup of the apps should be examined**, as these may present barriers to the IP CTS user community. All require downloading, setup, and some method of service eligibility verification before compensable use, in addition to basics like E911 registration and acknowledgement of 911 limitations disclaimers.

E. **Connection reliability in mobile platforms and applications varies in reliability or quality.** As with any mobile platform, users will be at the mercy of carrier data network availability and quality, geography, traffic loads, and even weather conditions. Few of these problems generally affect wireline based IP CTS availability.

F. **The proposed number of test subjects is not a statistically valid sample and will not be representative of the population to be served.** At just 10 participants, able to travel to MITRE's location, and "over 21 years old", the set of users is not a statistically valid sample of actual IP CTS users.

G. **The range, nature, length, and number of conversations are narrow and should be expanded to represent what IP CTS providers and users see in actual day-to-day service**. The scenarios listed in the Test Plan – a bank IVR and game with short yes-or-no responses – are not likely to produce data that will accurately predict results in day-to-day IP CTS implementations:
    - Challenge/Query responses found in business IVR systems don't call for (and, instead, limit) normal back-and-forth conversation interactions encountered by actual IPCTS users, in which the lengths, content, number of sentences and complexity of utterances will vary significantly.
    - A short word game is likely to be composed of few word phrases, unnatural wait-and-speak transactional behaviors, and brief overall speech that matches more closely the limited domain, directed speech models for which ASR-based "virtual assistants" are optimal rather than normal human conversation. Use of the NIST Switchboard corpus or, better yet, free-ranging live calls averaging four minutes length would provide more valid and useful results.

3. CaptionCall's recommendations for breadth and precision in tests

CaptionCall estimates that IP CTS providers currently handle around 300,000 calls per day and, during peak hours, 25,000 calls per hour at a rate of 400 to 500 per minute. System loads across the various providers likely total 2,000 or more concurrent sessions at any moment during daylight hours across the United States. Call lengths, while averaging around three to four minutes, can be as short as a few seconds to calls lasting for hours (e.g., between family members, friends, or businesses), with verbal content as diverse as the range of human

[9] https://store.google.com/us/product/pixel_phone?hl=en-US

[10] US Census Bureau, "Income and Poverty in the United States: 2015, Current Population Reports", Sept 2016, Page 6 Table 1

experiences, occupations, circumstances, and relationships. With this context in mind, CaptionCall has identified and recommends several additional elements be considered in the design of future tests to inform and guide the development of future IP CTS implementations and the regulations that guide service levels, minimum requirements and rate setting methods:

A. **Call content diversity.**
   a. **Speaker transitions during calls**, sometimes referred to as "Double-talk" or "Over-talk" as the give-and-take of conversation moves between speakers. On conventional PSTN lines, where a single conductor pair carries both the transmitted and received audio, the indefinite nature of speaker transitions presents significant challenges to human CAs as well as a challenge for ASR systems. When two people are speaking at once, with their speech mixed, the quality and accuracy of any speech-to-captioned text is affected.
   b. **Calling and called party audio volume.** Depending on age, disability, or just personal habit, users and those they converse with on the phone speak in a range of volumes, from shouting to whispers. ASR systems should be tested against this range to determine whether and what modifications should be made to ensure service reliability and quality.
   c. **Rate and length (of sections) of speech** (e.g. syllables or words per minute). As with the loudness of speech, people on phone calls may speak with amazing speed or extraordinary slowness, and with mid-sentence breaks and mid-word pauses. Most ASR systems now in use for domain-limited applications, such as "digital assistants" like Apple's Siri or Google Assistant, recognize fairly well and effectively present sentences or phrases containing one to eight words. However, when the speaker goes on at length, the transcript is presented as a single block of text without punctuation, capitalization or other sentence breaks – but including any mid-stream word recognizer errors – causing confusion to the assisted user. These behaviors on the part of people and systems should be part of test plans and evaluations.
   d. **Highly amplified sound.** Many users who experience hearing loss, desire and attempt to use every bit of audio they can. High levels of amplification create acoustic echo and automatic gain control management challenges, as well as making audio-stream artifacts that are not noticed at lower volumes, more apparent. Testing at high amplification levels is needed to verify any effects on ASR system accuracy.
   e. **Ambient or background sounds.** Calls to or from IP CTS users may contain significant ambient audio, whether from televisions, radios, appliances, or other people at the device location. Most ASR systems claim to filter non-speech sounds; however, the capability should be tested with some rigor.
   f. **Idioms and colloquialisms.** Common regional phrases, often combined with local accents, amount to dialects in speech. Idiomatic speech and strong dialects can cause colloquial challenges to ASR. Test scripts or recordings should include a sampling of such phrases and accents to verify performance.
   g. **Speaker age and gender.** In particular, the higher-pitched voices of young and female users can present challenges to automated systems ranging from noise filters and echo cancellers to ASR feature extraction engines. Test calls or recordings used to evaluate the efficacy of new technology implementations should be tested against these as well as middle and lower-pitched voices.
   h. **Accents.** Phase 1 and 2 tests included only one accented English instance and the Phase 3 plan indicates tests will continue with "various accents". CaptionCall encourages the CAMH to make available recordings, lists, or other samples of international accented speech to be used in tests.

B. **Connection and line conditions affecting audio signal quality**. Test models for any system considered for wide scale deployments should account for various equipment, wiring, and network conditions to ensure consistency of quality for users. Examples of tests to consider should include:
   a. **Analog PSTN phone lines.** While the Phase 3 test plan focuses on a high-end smart phone for the IP CTS user, they may be talking with people at locations using conventional phone equipment on ILEC PSTN or cable VoIP service lines. Tests that model this topology will verify performance when hybrid echo or other analog line artifacts occur.
   b. **Analog-to-analog.** As noted earlier, 76 percent of U.S. residents aged 70 and older have wireline phone service to their homes. Hybrid style devices (using POTS and Internet) now in common use for IP CTS should be specified and tested for use with new ASR systems.

C. **System Scale.** Perhaps the most fundamental requirement to validate the performance of any widely deployed system is to test it at the scale and performance levels handled by the system it is proposed to augment. Computational capacity could put limits on the extent to which ASR could be used in lieu of human CAs, even when all other technical issues are addressed. CaptionCall thus recommends that system test requirements be developed that at least approximate the following metrics for verified performance:
   a. **Call Volumes.** Systems for IP CTS national usage must be able to handle at least 300,000 calls per day. To handle peak daytime volumes, systems must be able to reliably support at minimum of 25,000 calls per hour.
   b. **Concurrent Session Handling.** The systems must support at least 2,000 concurrent sessions, and be able to expand by at least 50 percent to handle projected peak loads.
   c. **Rate of Session Creation.** The system must be able to successfully create, maintain, and generate required metadata at a rate of 500 sessions per minute.
   d. **Call Length.** Typical IP CTS calls are three to four minutes long. Test calls in Phase 1 and 2 all fell short of this average and the Test Plan does not specify the lengths of test calls planned for Phase 3. Given the behavior IP CTS providers see in actual calls by users, CaptionCall recommends that tests include calls at the average length for real calls, as well as shorter and longer lengths to model the range of actual user behavior.
   e. **Reconnection for interrupted sessions.** Occasionally, due to user's equipment, a network problem, ISP service failures, or user actions, IP CTS call sessions may be interrupted. IP CTS provider systems attempt to detect unintentionally interrupted calls and reconnect the user's device to resume captioning with only a brief break in the caption stream. Cloud-based ASR systems, and IP CTS deployments based on them, should provide similar functionality.

D. **Defining usability.** SUS and small group tests with few calls are inappropriate methods for accurately gauging the usability of captions delivered by IP CTS providers. An improvement would be:
   a. **The study should include a much larger number of participants or participant groups**, perhaps in various regions of the United States, to provide more statistically reliable assessments of typical IPCTS users.

4. **Process improvement and peer review**

While MITRE provided some informal opportunities for comment on Phase 1 and 2 test plans, in most cases recommendations from the providers were not implemented. The Phase 3 document marks the first time providers have been furnished plans and detailed prior work summaries in writing, enabling more considered and detailed comment on work completed and in progress. CaptionCall recommends a few additional steps to enable a more full understanding and informed critique of the work done to date and in planning, including:

A. Publishing written drafts of test plans for review and comment as early as practical, perhaps with deadlines for comments from providers or consumer groups wishing to make recommendations
B. Providing contact information for both test execution and usability assessment managers at MITRE or other consultancies used in the test program
C. Making available copies of recordings used or to be used in tests
D. Making available parametric files used with the SCLite scoring tool, with explanatory notes about configuration of the tool that may affect results
E. Making available to providers their specific individual raw test results (i.e., any single provider would see its own results, and not those of others)